dplyr and SQL

Randall Pruim

Big Data Ignite 2016

dplyr and SQL integration

`dplyr` provides SQL integration

- backends for several flavors of SQL exist
  - Postgres, MySQL, MariaDB, BigQuery
- others coming
  - MS SQL server

airlines data via MySQL (Smith College)

A data base that includes information on US flights since 1987 is available from Smith College

```
require(dplyr)
airlinesdb <-
  src_mysql(
    "airlines", host = "scidb.smith.edu",
    user = "mth292", password = "RememberPi")
airlinesdb
```

```
## src:  mysql 5.5.47-0ubuntu0.14.04.1 [mth292@scidb.smith.
## tbls: airports, carriers, flights, planes, summary, weat
```

Connecting to the tables

```r
# supressing a couple warnings about type conversion
Airports <- tbl(airlinesdb, "airports")
Carriers <- tbl(airlinesdb, "carriers")
Flights <- tbl(airlinesdb, "flights")
Planes <- tbl(airlinesdb, "planes")
Weather <- tbl(airlinesdb, "weather")
```

Taking a glimpse

```
Weather %>% glimpse()
```

```
## Observations: NA
## Variables: 15
## $ origin     <chr> "EWR", "EWR", "EWR", "EWR", "EWR", "E
## $ year       <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2
## $ month      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
## $ day        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
## $ hour       <int> 0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12
## $ temp       <dbl> 37.04, 37.04, 37.94, 37.94, 37.94, 39
## $ dewp       <dbl> 21.92, 21.92, 21.92, 23.00, 24.08, 26
## $ humid      <dbl> 53.97, 53.97, 52.09, 54.51, 57.04, 59
## $ wind_dir   <dbl> 230, 230, 230, 230, 240, 270, 250, 24
## $ wind_speed <dbl> 10.35702, 13.80936, 12.65858, 13.8093
## $ wind_gust  <dbl> 11.918651, 15.891535, 14.567241, 15.8
## $ precip     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ pressure   <dbl> 1013.9, 1013.0, 1012.6, 1012.7, 1012.
```

NY Weather only

Weather data is only for NY airports in 2013. Complete 24/7 data would include 8760 records.

```
Weather %>%
  group_by(origin, year) %>%
  summarise(num_records = n())
```

```
## Source:   query [?? x 3]
## Database: mysql 5.5.47-0ubuntu0.14.04.1 [mth292@scidb.sm
## Groups: origin
##
##   origin  year num_records
##    <chr> <dbl>       <dbl>
## 1    EWR  2013        8708
## 2    JFK  2013        8711
## 3    LGA  2013        8711
```

Airports

```
glimpse(Airports)
```

```
## Observations: NA
## Variables: 9

## Warning in .local(conn, statement, ...): Decimal MySQL c
## numeric

## Warning in .local(conn, statement, ...): Decimal MySQL c
## numeric

## $ faa     <chr> "04G", "06A", "06C", "06N", "09J", "0A9"
## $ name    <chr> "Lansdowne Airport", "Moton Field Munici
## $ lat     <dbl> 41.13047, 32.46057, 41.98934, 41.43191,
## $ lon     <dbl> -80.61958, -85.68003, -88.10124, -74.391
## $ alt     <int> 1044, 264, 801, 523, 11, 1593, 730, 492,
## $ tz      <int> -5, -6, -6, -5, -5, -5, -5, -5, -5, -8,
## $ dst     <chr> "A", "A", "A", "A", "A", "A", "A", "A",
```

GRR Airport

```r
Airports %>%
  filter(faa == "GRR")
```

```
## Source:   query [?? x 9]
## Database: mysql 5.5.47-0ubuntu0.14.04.1 [mth292@scidb.sm
##
##     faa               name      lat       lon   alt    t
##   <chr>              <chr>    <dbl>     <dbl> <int> <int
## 1   GRR Gerald R Ford Intl 42.88083 -85.52281   794    -
## # ... with 2 more variables: city <chr>, country <chr>
```

GRR Flights

```r
GRR2015 <-
  Flights %>%
  filter(year == 2015) %>%
  filter(origin == "GRR" | dest=="GRR") %>%
  collect()
```

```r
nrow(GRR2015)
```

```
## [1] 23540
```

Where do we go from here?

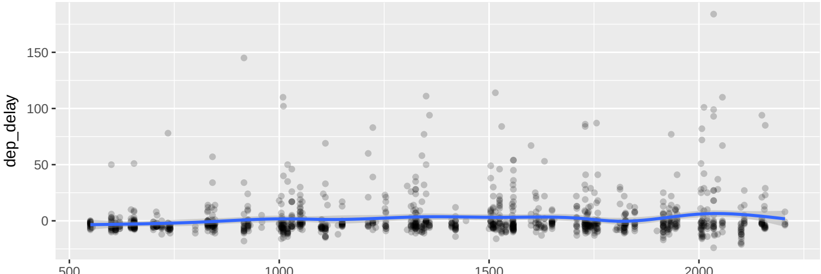
```
GRR2015 <-
  GRR2015 %>%
  mutate(date = lubridate::mdy(paste(month, day, year)))
GRR2015 %>%
  filter(origin == "GRR") %>%
  group_by(dest) %>%
  summarise(n = n()) %>%
  arrange(-n)
```

```
## # A tibble: 15 × 2
##     dest     n
##    <chr> <int>
## 1    ORD  3111
## 2    DTW  1572
## 3    ATL  1404
## 4    MSP  1177
## 5    DEN   854
```

Will my flight be on time?

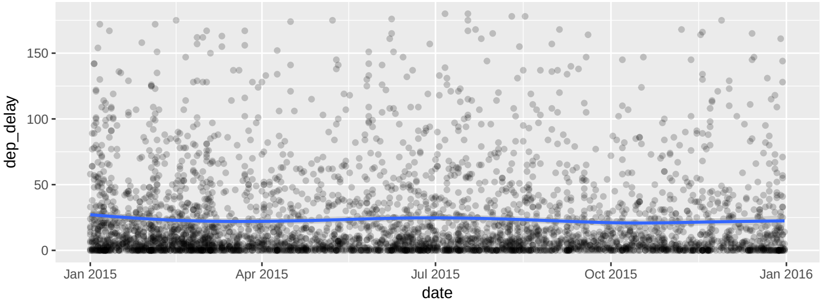
```
GRR2015 %>%
  filter((day > 20 & month == 9) |
         (day < 10 & month == 10)) %>%
  ggplot(aes(x = sched_dep_time, y = dep_delay)) +
  geom_point(alpha = 0.2) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam'
```



Will my flight be on time?

```
GRR2015 %>%
  filter(origin == "GRR") %>%
  ggplot(aes(x = date, y = dep_delay)) +
  geom_point(alpha = 0.2) +
  geom_smooth() +
  ylim(0, 180)
```


dep_delay
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
date